

# High throughput sequencing

Functional genomic data analysis: transcriptomics

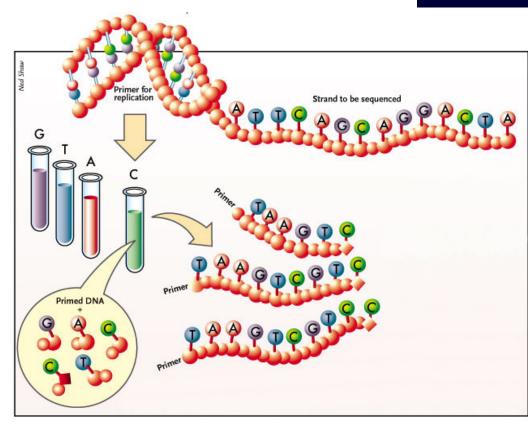


# First generation sequencing methods

# Sanger sequencing by synthesis



- Method discovered in 1977 by Frédérick Sanger (nobel price 1980).
- DNA polymerisation using a complementary primer.
- Elongation using thermostable DNA polymerase (PCR).
- Addition of 4 deoxynucleotides (dATP, dCTP, dGTP, dTTP) and low concentrations of one of four dideoxynucleotides (ddATP, ddCTP, ddGTP ou ddTTP).



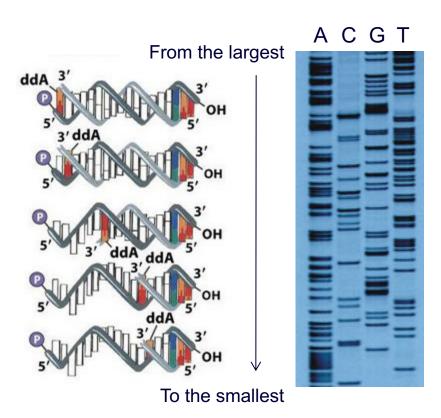
These ddNTP once incorporated in the newly synthesized DNA strand, block elongation. Synthesis termination is done by a statistical manner on each possible positions.

From The Scientist

## Sequence reading



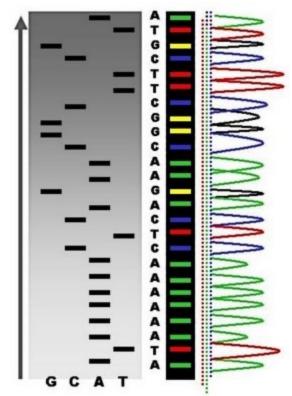
- We a get a mix of DNA fragments terminating at each position of the sequence.
- These fragments are then separated on a DNA polyacrylamide gel electrophoresis.
- Detection of synthesize fragments is done by the incorporation of labelling beacon in the DNA.
- At the origin this label was radioactive, attached either on the primer or on the dideoxynucleotide.
- Around 1 kb of DNA by run during 6-8 hours. One read by sample.

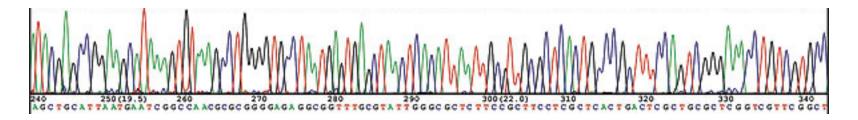


#### **Capillary sequencers**



- First version in the 90's thanks to the modification of the radioactive label by a fluorescent one.
- Using glass capillary of few micron diameters, on 30 to 50 cm long.
- The four nucleotides migrate in the same tube thanks to four different fluorescent dyes.
- **300 kb of DNA** by run during 3 hours. Several hundred sample at a time.

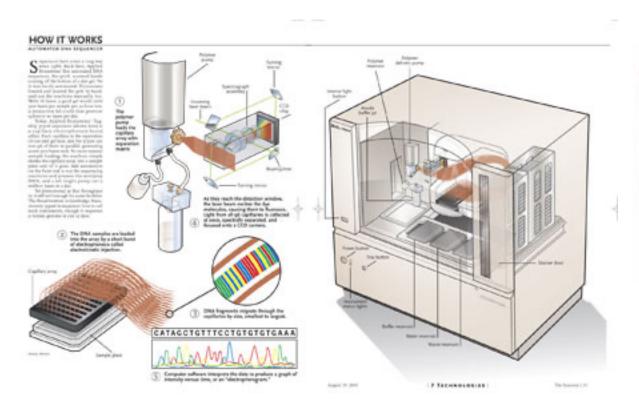




## **ABI 3730xl DNA Analyzer**



- = 96 parallel capillary (up to 50 cm) array. **768 samples**, **690 kb** DNA, **3 hours** run.
- At the Broad Institute (Cambridge, Massachusetts) 126 devices were able to sequence
   1 human genome in 12 days.







From The Scientist

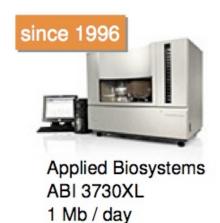


# Second generation sequencing: high throughput sequencing

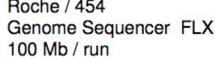
## The first technologies



Goal: to obtain a huge number of short reads.









Illumina / Solexa Genome Analyzer 2,000 Mb (2 Gb) / run



3,000 Mb (3 Gb) / run

SOLiD

#### **Genome Analyzer**



Available: January 2007, provider: Illumina

# illumına®

From: Clive Brown <clive.Brown@solexa.com>

Date: Sun, 20 Feb 2005 16:34:46 +0100

To: Nick McCooke <Nick.McCooke@solexa.com>, Tony Swerdlow <Harold.Swerdlow@solexa.com>, John Milt <Kevin.Hall@solexa.com>, Colin Barnes <Colin.Bar <Vincent.Smith@solexa.com>, Klaus Maisinger <Kla

Conversation: WE'VE DONE IT !!!!

Subject: WE'VE DONE IT !!!!



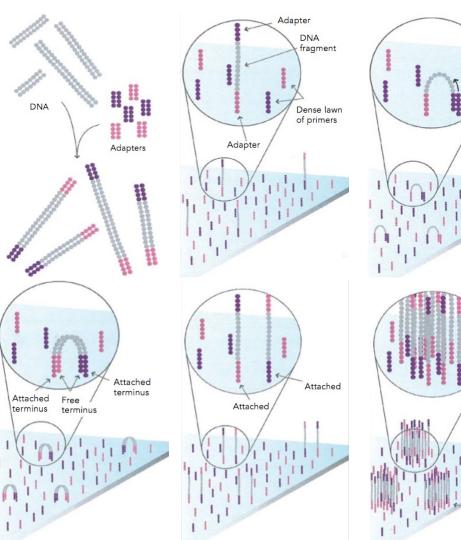
Tony Cox, Peta and I now agree - having looked at all of the PhiX174 data.

We have re-sequenced our first genome !!!!!!

# **DNA** library preparation



- Random DNA fragmentation and size selection.
- Ligation of adaptors.
- DNA denaturation.
- Hybridization of fragments onto the "flowcell" surface.
- Solid phase bridge PCR.



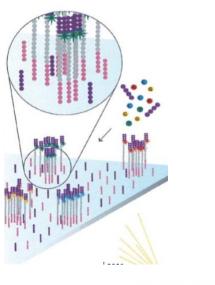
http://www.illumina.com/

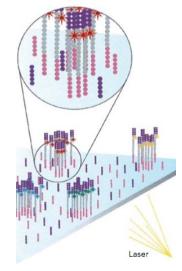
Clusters

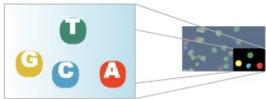
# Reversible terminator sequencing



- The four reversible terminators are added simultaneously.
- Laser scanning of the flowcell surface.
- Release of the blocking terminator.
- Sequencing cycles are repeated one base at a time.







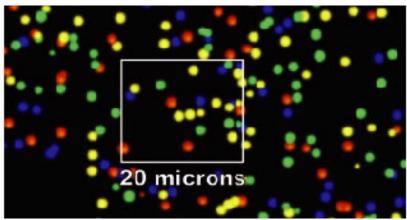


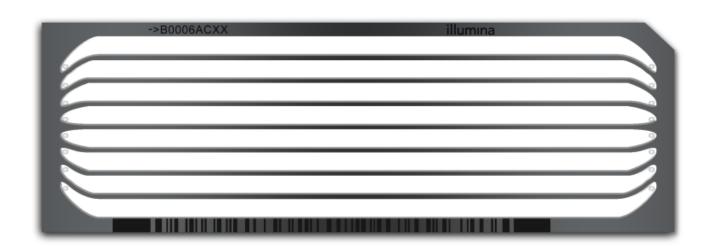
http://www.illumina.com/

# **Sequence analysis**



- Scanning at each position for all sequences (reads) in parallel.
- Most of the errors (99%) are sequencing errors.

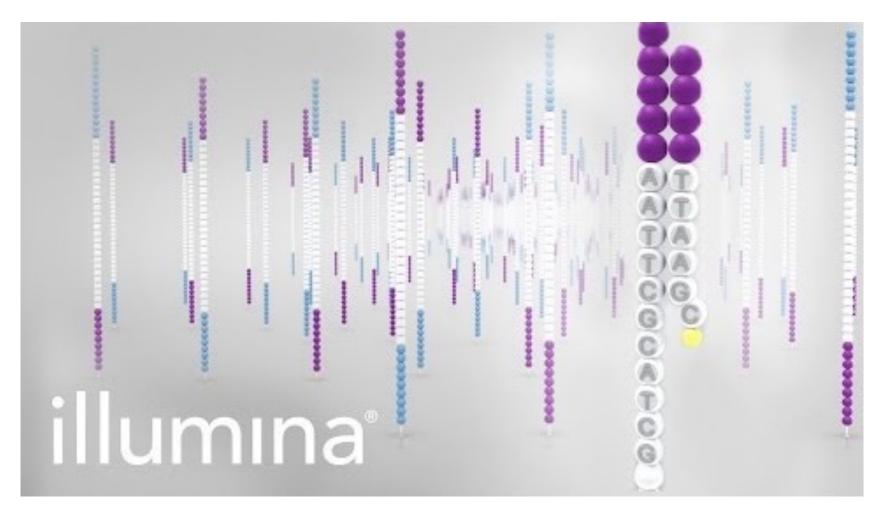




http://www.illumina.com/

# Illumina sequencing by synthesis video





https://www.youtube.com/watch?v=fCd6B5HRaZ8

# **Specifications of the latest Illumina sequencers**













	MiniSeq	MiSeq	NextSeq 550	NextSeq 2000	NovaSeq
Run Time	24 hours	56 hours	29 hours	2 days	44 hours
Read length (bp)	2x 150	2x 300	2x 150	2x 150	2x 150
Read number	50 10 <sup>6</sup>	50 10 <sup>6</sup>	800 106	1 10 <sup>9</sup>	10 10 <sup>9</sup>
Ouput	7.5 Gb	15 Gb	120 Gb	300 Gb	3,000 Gb
Throughput	7 Gb/day	7 Gb/day	100 Gb/day	150 Gb/day	1,500 Gb/day

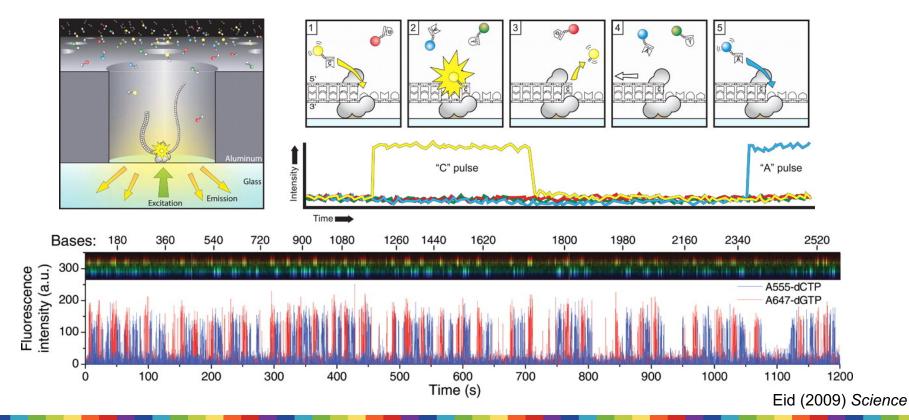


# The third generation

# Real time sequencing



- Real time sequencing on single molecule thanks to RNA polymerase immobilisation in wells.
- Each base incorporation is measure in real time with a CCD camera under the bottom of the plate.



# **Sequel specifications**



- 1 000 000 reads/ SMRT cell;
- From 1 to 45 kb (Avg 20 kb);
- 10 Gb by SMRT cell;
- Run duration = 6 hours.



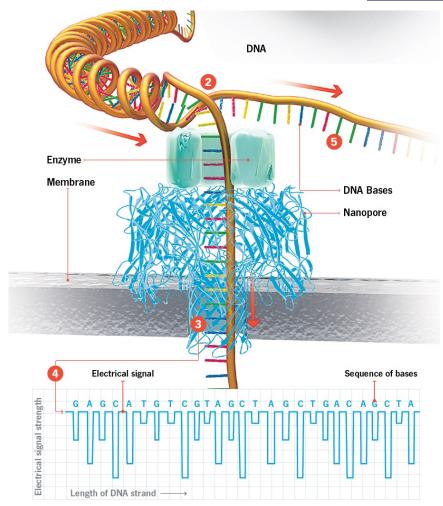


# Nanopore technology



- Single molecule detection system by passing single strand DNA through a nanometric pore.
- Base to base analysis using electric properties of the nanopore.

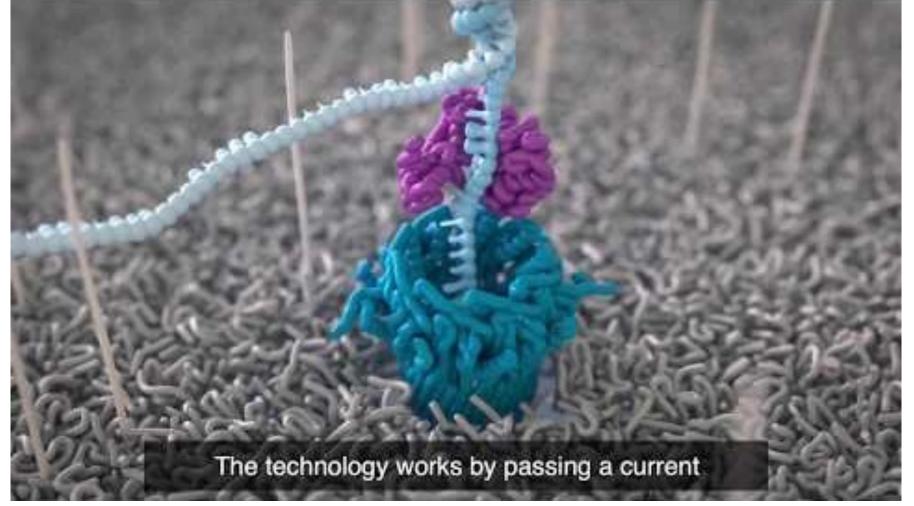
- DNA size sequencing of kilobases
- No limitation on the acid nucleic type to be detected (DNA or RNA).
- No amplification.



Greenwood (2013) Popular Science

# Oxford nanopore technologies sequencing video



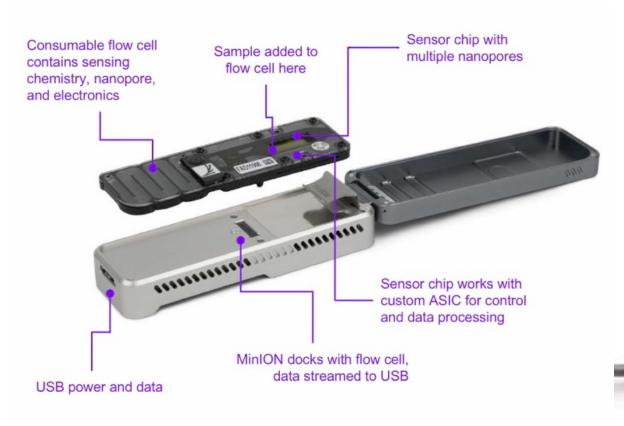


https://www.youtube.com/watch?v=RcP85JHLmnI

#### The MiniON flow cell



- 1 flow cell = 1 membrane with 514 nanopores.
- Single molecule sequencing up to 100 kb during up to 48 hours.





@NanoporeConf



# **Applications**

# They cover a lot of already existing techniques



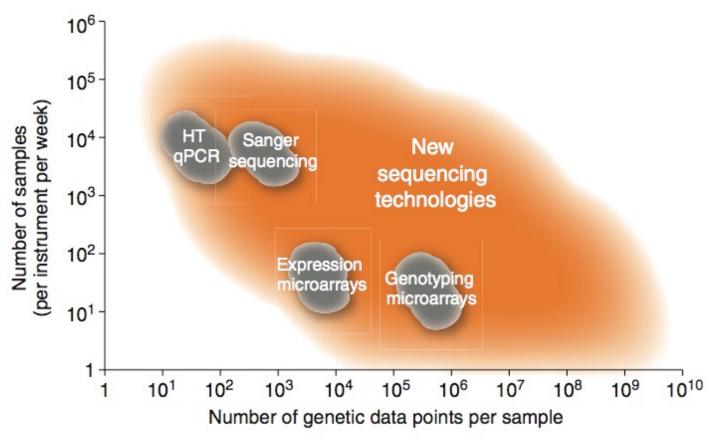


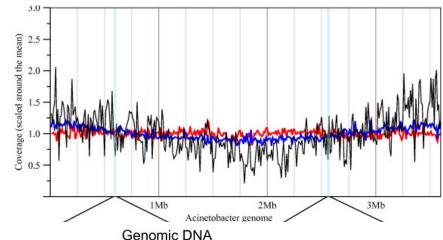
Figure 2 Relative sample and data throughputs for different nucleic acid detection and sequencing technologies. A rough estimate of the number of samples that can be run on a single instrument in one week with the resultant data points is shown on a logarithmic scale for different technologies. This is intended for comparative scale and is not exact.

Kahvejian et al. (2008) Nat. Biotech.

#### De novo sequencing



- Quicker and cheaper sequencing than Sanger.
- But reads are smaller.
- Combination of different methods allow to obtain better quality sequencing drafts.
  - => Combining 454 and Illumina.
- Low error rate and homogenous coverage due to no cloning biases compared to the Sanger method.
- Errors are not same with the two high throughput sequencing methods.



Roche-454 sequenced paired-end library to a ~7x fragment size coverage (for 3 kb fragments)

Add 454 unpaired data to a final 25x coverage

Newbler Assembly

Correct errors with ~50x Solexa-Illumina short reads data

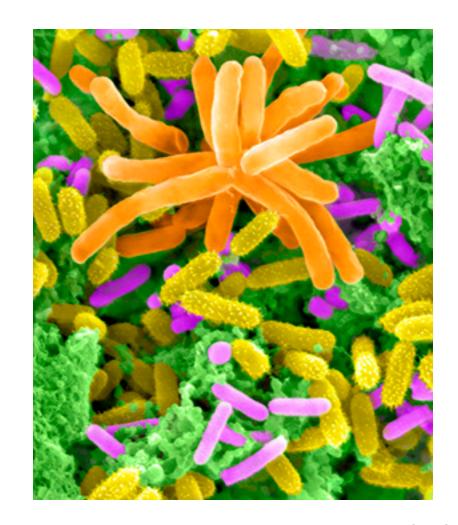
High quality draft (<10<sup>-4</sup> error rate)

Aury et al. (2008) BMC Genomics

# Resequencing applications



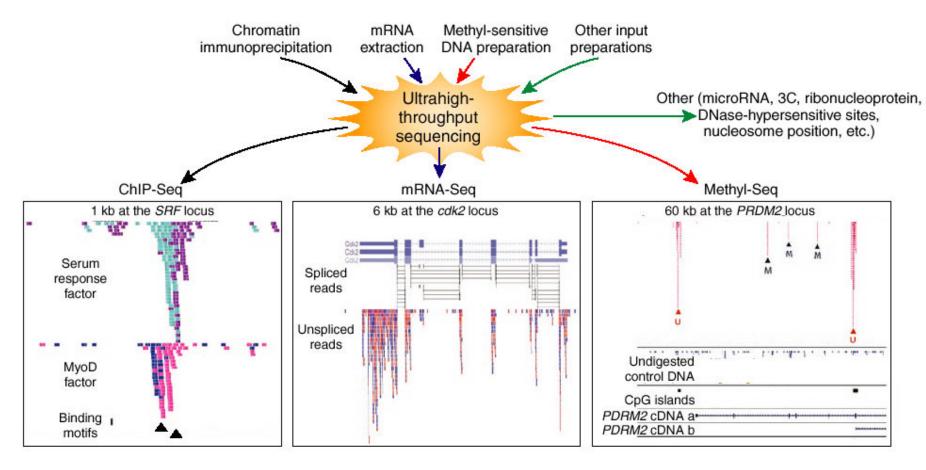
- Goal: analyze various genomes compared to a reference one.
- Search for polymorphisms and structural variants in populations, mutation identification in biotechnology, organism evolution analyses, cell differentiation along time, ancient DNA discovery...
- Metagenomics: genome characterisation in samples.
- A wide range of applications: characterise pathogen micro-organisms in patient tissues, definition of the species found in environment samples, understand species evolution...



From JGI DOE

# **Functional genomic applications**





Wold et al. (2008) Nat. Methods

# High throuput sequencing outcome



#### **Avantages**

- No cloning steps in bacteria
  - no more bias;
  - library construction easier.
- Each sequence come from a unique molecule:
  - quantification;
  - larger dynamic range.
- High resolution for a wide range of different applications.
- Huge improvements in term of speed and cost compared to the former Sanger method.

#### **Drawbacks**

- Sequences are shorter
  - compare to Sanger;
  - new base calling parameters;
  - new bioinformatics analyses.
- The amount of results generated is a problem for data management:
  - several To by run;
  - CPU usage;
  - what can we store?
- A costly still evolving technology.
- **=** Library preparation is a limiting step.